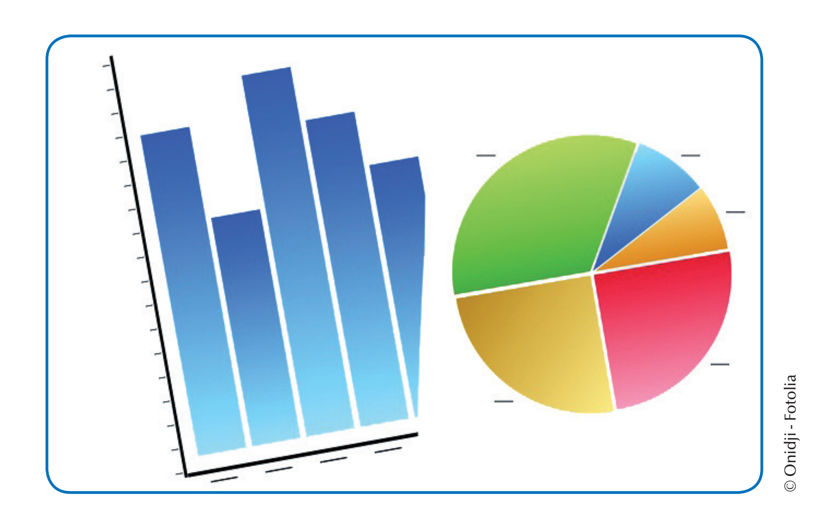
Etat des lieux sur « le test »

Qu'appelle-t-on un test ? Qu'est-ce qu'évaluer ? Du test au testeur... Comment franchir le pas ?

Simon Marin-Curtoud[1], Thierry Rousseau[2], Peggy Gatignol[3]



Le concept d'évaluation incite généralement à évoquer la figure d'Alfred Binet, « créateur » du premier test psychologique : **l'échelle métrique de l'intelligence** *(Binet et Simon, 1905)*, dont le propos était d'identifier les enfants pouvant tirer bénéfice d'un enseignement spécialisé. Cette évocation trouve également sa justification dans le fait que l'évaluation d'Alfred Binet s'inscrivait dans le cadre des évaluations diagnostiques, à visée de remédiations. En effet, il est d'usage de distinguer les évaluations en fonction de leurs objectifs :

- l'évaluation sommative, visant à faire un bilan des connaissances,
- l'évaluation certificative, amenant à la production d'un diplôme,
- l'évaluation formative, permettant de situer l'apprenant dans un parcours d'apprentissage, voisine de l'évaluation diagnostique, les objectifs de cette dernière étant de détecter les causes d'un déficit en vue d'y remédier, *Vrignaud, (2004)*.

[1] Orthophoniste, 49 rue de Rivoli, 75001 Paris
[2] Orthophoniste, Dr en psychologie, 11 avenue Joël Le Theule 72303 Sablé/Sarthe
[3] Orthophoniste, Dr en neurosciences, pôle tête et cou, hôpital Pitié Salpêtrière 75013 Paris

Définition

Définir le concept de test n'est pas chose aisée. Pour Marie-Noëlle Metz-Lutz (1988), le test est comme *« une épreuve impliquant une tâche définie pour tous les sujets et comportant une technique précise pour l'appréciation des succès et des échecs »*.
Pour sa part, Pierre Pichot (1997) : *« On appelle test mental une situation expérimentale standardisée servant de stimulus à un comportement »*. Ce dernier étant destiné à être comparé statistiquement à celui d'autres sujets placés dans les mêmes conditions expérimentales, formant un groupe de référence par rapport auquel le sujet testé pourra être comparé.

La Société Française de Psychologie, à travers son adaptation française des ***Recommandations Internationales sur l'Utilisation des Tests*** *(Vrignaud et coll., 2003)* considère que *« toute tentative pour fournir une définition précise d'un test [...] en tant que processus échouera vraisemblablement parce qu'elle risque d'exclure certaines procédures qui devraient en faire partie, et d'en inclure d'autres qui devraient en être exclues »*.
Face à cet obstacle, *les Recommandations* optent pour une série de propositions visant à organiser le domaine visé. Nous en soulignerons quatre :

- la passation de tests comprend des procédures permettant la mesure des comportements normaux ou pathologiques, voire des dysfonctionnements,
- les procédures de passation de tests sont habituellement construites pour être administrées selon des conditions soigneusement contrôlées ou standardisées, qui incluent des protocoles cotés de manière systématique,
- ces procédures fournissent des mesures de la performance et amènent à tirer des inférences à partir d'échantillons du comportement,
- elles comprennent également des procédures qui peuvent aboutir à catégoriser ou à classer les personnes.

Par ailleurs, les *Recommandations* stipulent que toute procédure se réclamant de l'appellation de «test» doit pouvoir s'appuyer sur des constats de fidélité et de validité en relation avec les objectifs poursuivis et fournir les preuves à l'appui des inférences tirées des scores aux épreuves considérées.



© Hannes Eichinger - Fotolia

Huteau et Lautrey (1997) précisent qu'un test est un dispositif d'observation des individus qui présente quatre propriétés :
- il est standardisé,
- il permet de situer la conduite de chaque sujet dans un groupe de référence,
- le degré de précision des mesures qu'il permet est évalué (fidélité),
- la signification théorique ou pratique de ces mesures est précisée (validité).

Face à ces notions de standardisation, de fidélité et de validité, il est primordial de connaître et maîtriser la méthodologie des tests.

Méthodologie des tests

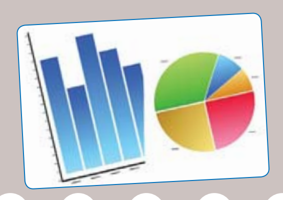
La méthodologie psychométrique, issue de la méthode expérimentale, intègre en son champ l'étude des différences individuelles. Elle se caractérise par une standardisation des conditions d'observation, une réflexion approfondie sur la mesure et sa signification et la formalisation des notions d'erreur de mesure et d'erreur de pronostic.

Standardisation

Selon *Huteau et Lautrey (1997 (1999),* la standardisation d'une procédure d'observation constitue la caractéristique essentielle des tests. Elle vise, sinon à éliminer totalement, du moins à réduire les biais dus à l'observateur. On parle d'« objectivité » dans une observation lorsque plusieurs observateurs indépendants décrivent de manière identique la conduite d'un sujet observé. Il convient cependant de garder à l'esprit que l'objectivité d'une observation ne garantit en rien le risque d'erreur systématique commune à tous les observateurs ni n'atteste, à priori, de la pertinence d'une observation.

Observation automatisée

On peut considérer deux types d'observation automatisée : le testing collectif et le testing informatisé.
Le testing collectif implique que les sujets soient en groupe, les consignes et problèmes soient présentés par écrit, et que quelques indications orales soient données à l'ensemble du groupe. Dans cette configuration, les interactions entre les sujets et celui qui administre le test sont réduites au minimum.
En général, le questionnaire sous forme de QCM permet une correction automatisée.
L'informatisation des tests s'est répandue à la faveur du développement de la micro-informatique dans les années 70.
Avec l'informatisation, l'interaction sujet-administrateur de test peut totalement disparaître puisque la présence d'un applicateur semble devenir superflue.
Pour *Huteau et Lautrey (1999),* la standardisation, dans ces modes d'observation, est parfaite, tous les biais liés à la variabilité des observateurs ayant été écartés. Il convient alors de s'interroger sur la nature des observations réalisées et d'éventuels biais systématiques.



Par exemple, dans quelle mesure les différenciations interindividuelles établies dépendent-elles du mode d'observation retenu ? A condition de prendre un certain nombre de précautions, il semble que l'effet du mode de questionnement soit faible.

Limites de l'observation standardisée

Si l'observation standardisée offre l'avantage de réduire considérablement les biais dus à l'observateur, elle constitue aussi une limite à ce qui est observable : toute nouveauté, n'entrant pas dans le cadre pré-établi, est inutilisable. Ainsi, les méthodes standardisées sont-elles plus adaptées aux domaines bien défrichés qu'aux explorations de conduites peu connues. La réserve communément formulées à l'encontre des méthodes standardisées repose sur leur supposée incapacité à tenir compte du contexte et à appréhender une réalité sous-jacente à une conduite manifeste.

La mesure

Rappelant qu'au sens très général, mesurer, c'est attribuer des nombres aux choses, *Huteau et Lautrey, (1999), citant Reuchkin (1970)*, précisent que *« pour que les propriétés des nombres puissent être appliquées aux choses, il est indispensable de fonder les correspondances entre ces propriétés des nombres et les propriétés des choses. »*

Niveaux de mesure

Depuis les travaux psychophysiques de *Stevens (1951)*, il est communément admis de distinguer quatre niveaux hiérarchisés de mesure, ou quatre types d'échelles de mesure.

- ***Les échelles nominales*** : utilisables dans le cas d'observations regroupables en classes d'équivalence, chacune de ces classes pouvant être désignée par un nombre. Les nombres n'ont ici que la propriété d'être des symboles distincts, qu'il n'y a par conséquent aucun sens à ordonner ou ajouter. Ce niveau de mesure (relativement faible) incite à parler de mesure qualitative et permet le traitement statistique.
- ***Les échelles ordinales :*** utiles lorsque l'on peut établir un ordre entre les classes et montrer que les relations inter-classes sont antisymétriques (si A>B, alors B>A est impossible) et transitives (si A>B et B>C, alors A>C). Les nombres désignant les classes deviennent alors des symboles ordonnés.
- ***Les échelles d'intervalles :*** adaptées à un traitement visant à définir des distances entre les classes, un intervalle-unité permettant de définir de nouveaux intervalles (concaténation). Les nombres acquièrent alors de nouvelles propriétés. L'unité étant conventionnelle et l'origine arbitraire, toutes les transformations numériques de la forme $y = ax + b$ sont permises. Il est donc possible à ce niveau de calculer une moyenne, une variance ou un écart-type.
- ***Les échelles de rapports :*** utilisées si l'on peut non seulement définir des intervalles entre les classes, mais aussi une origine, ou montrer que le rapport numérique entre deux classes est égal au rapport numérique entre 2 autres classes.

Etalonnages

Pierre Pichot (1997) nous rappelle que *« le test est un instrument de mesure constitué d'éléments, ou items, dont l'ensemble constitue une échelle. La cotation vise à transformer la réponse à un item en une valeur numérique suivant des règles pré-établies. La somme des notes obtenues aux items constituant l'échelle est la note brute à l'échelle. Or cette note brute n'acquiert une signification et ne devient mesure que lorsqu'elle est rapportée à un étalon. »*

Les étalonnages sont donc des systèmes de catégories ordonnées dans lesquelles il est possible de ventiler tous les sujets d'un groupe de référence.

Il existe alors deux grandes catégories d'étalonnages :

- les quantilages : catégories ordonnées de mêmes effectifs,
- les échelles normalisées : partition, selon certaines règles, d'une distribution normale (distribution théorique de *Laplace-Gauss*).

La méthode des quantiles consiste à regrouper les notes brutes obtenues de manière à obtenir des catégories à effectifs identiques. Cette méthode, bien que facile à construire, est parfois critiquée pour son manque de différenciation des sujets se



© Karen Roach - Fotolia

trouvant aux extrémités de la distribution et son excès de distinction des sujets occupant le centre de la distribution.

C'est pour pallier cet inconvénient que l'on utilise parfois des échelles normalisées (écart-type) : dans ce type d'étalonnage, les catégories sont toujours définies par des effectifs, mais ceux-ci ne sont plus égaux. Leurs limites sont déterminées de façon à ce qu'en considérant qu'elles définissent des intervalles égaux, il soit possible de reconstituer une distribution proche de la distribution normale.

Tests à références critérielles

Le test à référence critérielle permet de situer la performance d'un sujet par rapport à un univers de contenu (connaissances, compétences) et se distingue du test à référence normative qui lui, permet de situer la performance du sujet par rapport à un groupe.

La définition des dimensions : 3 modèles de mesure

Rappelons qu'un test est un ensemble d'items donnant chacun lieu à un score et que ces résultats sont additionnés pour obtenir un score d'échelle.

Pour *Huteau et Lautrey (1999)*, cette pratique, si elle est doublement justifiée (elle fournit une bonne différenciation des individus et permet de neutraliser certaines erreurs de mesure), doit nous amener à nous interroger sur la pertinence de l'opération consistant à additionner des scores partiels, ou, en d'autres termes, si tous les items contribuent bien à la mesure d'une même dimension, ou encore, si tous les items constituent bien un ensemble homogène. Les auteurs nous proposent trois modèles dits *« modèles de mesure »* visant à répondre à ces questions. Ces modèles sont présentés, ainsi que d'autres, dans l'ouvrage de *Dickes et coll.*, (1994).

Analyse d'items et corrélation item-test

Il s'agit de la méthode de construction de test la plus répandue et la plus conviviale.
Cette méthode consiste à partir d'une définition conceptuelle d'une dimension à évaluer puis à élaborer un ensemble d'items de difficulté graduée impliquant cette dimension. Ces items seront ensuite soumis à un ou plusieurs groupes de sujets afin que ne soient conservés que ceux des items qui permettent une bonne différenciation des individus et qui constituent un ensemble homogène. Il convient donc de disposer au départ d'un nombre d'items nettement supérieur à ce que l'on souhaite pour le test final. Cette démarche ne fait pas l'économie d'un questionnement sur l'éventuelle remise en cause, par la sélection d'items opérée, de la définition initiale de la dimension.
Chaque item est caractérisé par un *« indice de difficulté »,* qui n'est autre que la fréquence de réussite à cet item, déterminant son pouvoir de différenciation des individus.
Un item a un pouvoir de différenciation maximum lorsque sa fréquence de réussite est de 50 %, il est nul lorsque cette fréquence s'approche de 0 % (personne ne réussit) ou de 100 % (tout le monde réussit).
Il est généralement admis que les items dont la fréquence de réussite est comprise entre 20 % et 80 % peuvent être retenus.
D'une grande simplicité, cet indice est néanmoins dépendant du groupe de sujets considérés. Les *« modèles de réponse à l'item »* viennent minorer cet inconvénient.
Par ailleurs, chaque item peut être caractérisé par un *« indice de discrimination »,* permettant de distinguer les items selon leur contribution au score final. Cet indice de discrimination est le coefficient de corrélation entre l'item et le score au test. Il est élevé si les individus qui réussissent l'item ont un score élevé au test, faible dans le cas où il y a peu de rapport entre la réussite à l'item et le score au test.
Plus l'indice de corrélation item-test est élevé, plus l'homogénéité du test est forte. Cependant, ce cas de figure est fréquent avec les tests dans lesquels les items sont très proches les uns des autres et n'évaluent que des secteurs très étroits de la conduite. En conséquence, le seuil d'élimination des items en fonction de leur indice de corrélation item-test est généralement assez bas (autour de 30%).
Cette procédure correspond à un modèle de mesure appelé *« la théorie classique du score vrai ».* Son postulat est qu'il existe une variable latente au long de laquelle on peut ordonner les sujets selon leurs notes « vraies », c'est-à-dire indemnes d'erreur de mesure. Cependant le test, variable observable, ne correspond pas exactement à cette variable latente, du fait des erreurs de mesure inhérentes au choix particulier des items.
Ces erreurs de mesure peuvent néanmoins être évaluées.


© Dominique Luzy - Fotolia

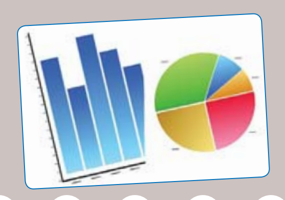
Modèles de réponse à l'item

C'est dans les années 50, partant de l'analyse d'items et à la faveur des avancées en micro-informatique, que sera développé la forme des modèles de réponse à l'item, mais ce n'est que récemment que ces modèles ont connu leur développement, à la faveur des avancées en micro-informatique, *Vrignaud, (1996).*
Le modèle précédent visait à définir des courbes caractéristiques d'items (relation entre la réussite à un item et la réussite au test). Ces courbes étaient construites à partir d'observations.
Dans les modèles de réponse à l'item, les courbes sont définies à priori et représentent la relation entre la probabilité pour un sujet de réussir un item (et non plus la fréquence de réussite dans un groupe) et sa position sur une variable latente (et non plus son score sur une variable observable).

Analyse factorielle

A la différence des méthodes précédentes, il s'agit d'une méthode d'analyse multidimensionnelle.
Elle s'appuie sur deux postulats :
- lorsque plusieurs variables sont en corrélation, elles sont sous la dépendance d'un ou plusieurs facteurs communs de variation. Ces facteurs sont donc des abstractions mathématiques, qui ont le statut de variables latentes,
- les scores dans les variables observées sont des combinaisons linéaires des scores dans les variables sources.

Afin d'en approcher le principe, *Huteau et Lautrey (1999)* proposent une analogie : imaginons un examen comportant 4 matières, chacune des matières ayant des coefficients différents selon la section envisagée par le candidat. A notes identiques, un candidat n'aura donc pas la même moyenne suivant la section présentée.
La moyenne d'un sujet est donc une variable issue de la combinaison linéaire de plusieurs variables-sources (les notes à chaque épreuve) dont le poids dans la combinaison est fonction des coefficients qui leurs sont affectés. L'analyse factorielle consiste en la résolution d'équations équivalentes à celle-ci à ceci



près que le problème est posé à l'inverse : on dispose ici de la moyenne obtenue et les inconnues sont constituées par les scores dans les variables-sources et les coefficients de pondération.
Les raisonnements mathématiques impliqués dans ces calculs dépassent le cadre de notre propos et se trouvent détaillés dans différentes publications de langue française, *Bacher et Reuchlin, (1989) ; Cibois, (1983) ; Reuchlin, (1964, 1970) ou de langue anglaise, Kim et Mueller, (1978).*

Qualités métrologiques des tests

Les fidélités

Tout test doit être fidèle, dans les deux sens du terme, c'est à dire présenter une bonne stabilité temporelle et une bonne consistance interne.
Le coefficient consistance interne est une mesure de la fidélité. Il se calcule en utilisant *l'alpha de Cronbach,* basé sur le rapport entre la part de variance de chaque item et la variance de l'ensemble des scores de l'échelle.
Pour *Beech et Harding (1994),* le coefficient de consistance interne du test, d'une valeur minimale de 0.70 devrait être obtenu à partir d'un échantillon de 100 sujets au moins.
La corrélation des notes obtenues au même test administré à deux reprises représente la fidélité test-retest. Idéalement égale à 1.00, cette fidélité est jugée satisfaisante à partir de 0.70. Pour les auteurs, pour être pris en compte, ce coefficient devrait toujours être calculé sur les performances d'un échantillon de 100 individus au moins et l'intervalle entre les deux administrations devrait se situer entre un et trois mois.

Evaluation des erreurs de mesure

On en distingue deux types :
• des erreurs aléatoires, variables d'une observation à l'autre et imprévisibles au niveau de contrôle des observations choisi,
• des erreurs systématiques se manifestant de la même manière d'une observation à l'autre.
La théorie de la fidélité ne traite que les erreurs aléatoires. Cette théorie s'applique à des mesures ayant les propriétés des échelles d'intervalles.
C'est à la faveur des variations de la mesure lorsqu'elle est répétée que l'on prend conscience des erreurs de mesure.
La théorie de la fidélité postule que les mesures résultant de la réplication se distribuent normalement (l'erreur est aléatoire). Cette erreur a donc autant de chances de se manifester en positif qu'en négatif. La théorie considère qu'il est possible de décomposer toute mesure observée en deux parties indépendantes : une « mesure vraie » non observable à laquelle vient s'ajouter une erreur aléatoire.
Une première estimation de l'importance de l'erreur aléatoire est fournie par la dispersion des mesures observées répétées sur un même sujet. On appelle variance d'erreur la variance de cette distribution intra-individuelle et écart-type l'erreur standard de mesure. Plus cette distribution est dispersée, plus l'erreur est grande, plus la fidélité est faible.
Cependant, cette estimation dépend de l'unité de mesure choisie et n'a pas la même signification selon que la dispersion interindividuelle des notes observées est forte ou faible. C'est pourquoi l'on définit généralement la fidélité par le coefficient de fidélité r = variance des notes vraies / variances des notes observées (coefficient de généralisabilité).
De la même façon que nous avions et puisque la mesure vraie et l'erreur sont indépendantes :
• mesure observée = mesure vraie + erreur aléatoire
nous avons également :
• variance des notes observées = variance des notes vraies + variance d'erreur
Donc nous avons :
• variance des notes vraies = variances des notes observées – variance d'erreur.
En l'absence d'erreur, $r = 1$. Plus la part d'erreur augmente dans la mesure observée, plus r diminue.
On sait que l'une des difficultés de l'observation en psychologie est qu'elle modifie le sujet observé, ce qui nous empêche de procéder à des estimations directes de la variance d'erreur pour chaque sujet car cela supposerait de nombreuses répétitions de la mesure sur le même sujet.
Pour pallier cet inconvénient, on s'en tient généralement à une répétition de la mesure, la variance d'erreur intra-individuelle étant alors calculée sur l'ensemble des sujets.
Dans la pratique, il est fréquent de s'en remettre au coefficient de corrélation test-retest pour définir le coefficient de fidélité.

Sources d'erreur

On peut distinguer trois sources d'erreurs, relevant de trois modalités de répétition de la mesure :
• le sujet est placé dans la même situation à des moments différents ; les erreurs d'observation proviennent des facteurs associés au moment de l'observation, les coefficients de fidélité sont des coefficients de stabilité ou de constance,
• le sujet se voit proposer des épreuves différentes dans leur contenu, mais censées mesurer la même chose. Les erreurs d'observation proviennent alors de la spécificité des tâches proposées. Les coefficients de fidélité calculés sont des coefficients d'homogénéité,
• le sujet ne passe qu'une seule fois le test, la cotation et l'évaluation étant proposées à plusieurs observateurs. La source d'erreur repose donc sur l'observateur. Le coefficient calculé est le coefficient de fidélité inter-observateurs. C'est dans le but de maximiser cette fidélité que les conditions d'observation sont standardisées.

Les validités

« Un test est dit valide lorsqu'il permet d'atteindre de manière satisfaisante les objectifs que le constructeur ou l'utilisateur ont choisis. ». Huteau et Lautrey, (1999).
On envisage donc autant de types de validités que de catégories d'objectifs et l'on peut en dégager trois principales :

La validité de contenu

Etiquette sous laquelle sont rassemblés les tests constituant un échantillon représentatif de tâches ou de conduites. Ce sont par exemples les tests issus des programmes éducatifs. En effet, les programmes éducatifs visent à doter les individus d'un ensemble de propriétés (connaissances, compétences spécifiques ou plus générales). Les tests présentant un ensemble d'items représentatifs d'une catégorie d'objectifs éducatifs sont dits

avoir une bonne validité de contenu. *Walton et Bartram, (1994)* distinguent cette validité de contenu d'une validité apparente en ce que la première est estimée par un groupe de professionnels alors que la seconde est estimée par l'individu testé.

La validité empirique (ou critérielle)

La validité empirique désigne une forte corrélation entre un test et un critère (ou variable). Ici, l'objectif n'est plus de savoir quel trait sous-jacent est mesuré par le test, mais de savoir si le test est un prédicteur correct du critère. Le domaine dans lequel la validité empirique s'exprime avec le plus de force est certainement celui du recrutement : pour une entreprise capable de définir précisément ce qu'elle entend par « réussite » il sera possible d'examiner les relations entre ces mesures de réussite (critères) et les scores à des tests (prédicteurs).

La validité théorique

La validité théorique (ou validité hypothético-déductive, ou validité de construction), fait référence aux tests construits pour mesurer des traits hypothétiques. L'intelligence n'a pas d'existence physique, non plus que l'aptitude spatiale, pourtant, des tests ont été construits pour tenter d'évaluer ces qualités. Ceci fait dire à *Walton et Bartram (1994)* que *« la validité de construit renvoie à ce que nous savons et à ce que nous comprenons de la signification du score fourni par un test. Cette connaissance peut s'élaborer de manière inductive ou déductive. Cependant [...] quelle que soit la méthode par laquelle elle a été acquise, [elle] doit permettre [...] de prédire des comportements des sujets en situation réelle ».*

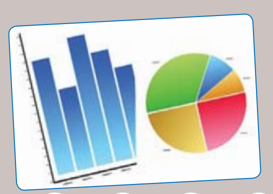
Pour les auteurs, validité d'un test ne se résume pas à sa propension à effectivement mesurer ce qu'il est censé mesurer. Les auteurs préfèrent adopter une autre définition de la validité : *« la validité renvoie à la pertinence et à la possibilité de justifier les affirmations que l'on peut faire à partir des scores à un test, elle concerne également les éléments dont on dispose pour justifier les inférences que l'on peut faire à partir des scores à un test. ».*

Informatisation des tests : quelle est la pertinence et quelles sont les limites des outils d'évaluation informatisés ?

Rappelons qu'un test est un ensemble d'items donnant chacun lieu à un score et que l'ensemble de ces scores est additionné pour obtenir un score d'échelle.
Pour *Huteau et Lautrey (1999)*, cette pratique est doublement justifiée, car elle fournit une bonne différenciation des individus et permet de neutraliser certaines erreurs de mesure. Cependant, il est nécessaire de s'interroger sur la pertinence de l'opération consistant à additionner des scores partiels. La question est de savoir si tous les items contribuent bien à mesurer la même dimension, ou encore, s'ils constituent bien un ensemble homogène. Cette méthode part d'une définition conceptuelle de la dimension à évaluer, puis sélectionne un ensemble d'items de difficulté graduée impliquant cette dimension. Ces items sont ensuite soumis à un ou plusieurs groupes de sujets, afin de ne conserver que ceux qui permettent une bonne différenciation des individus et constituent un ensemble homogène. Chaque item est caractérisé par un « indice de difficulté », qui n'est autre que la fréquence de réussite à cet item, déterminant son pouvoir de différenciation des individus. Un item a un pouvoir de différenciation maximum lorsque sa fréquence de réussite est de 50 %, il est nul lorsque cette fréquence s'approche de 0 % (personne ne réussit) ou de 100 % (tout le monde réussit). En conséquence, le seuil d'élimination des items en fonction de leur indice de corrélation item-test est généralement assez bas (autour de 30 %).

L'informatisation des tests s'est développée à la faveur de l'essor de la micro-informatique dans les années 1970. Comme le note *Bonin, (2003)*, l'identification de facteurs qui contribuent *« aux variations des latences d'initialisation est importante car elle autorise ensuite la détermination, au sein d'une architecture fonctionnelle, du ou des locus(i) d'impact de cette variable et le ou les mécanismes qui en sous-tend(ent) l'effet. ».*

L'informatisation des tests permettant de maîtriser la variable « temps de traitement d'une tâche », différentes épreuves ont été élaborées et validées pour tenter d'évaluer ce paramètre.
Les évaluations assistées par ordinateur permettent d'étudier l'évolution parallèle de la précision de la réponse et des temps de réponse, *French (1994) ; Le Gall et Allain, (2001)*. Le clinicien peut alors observer d'éventuelles dissociations entre une performance rapide accompagnée d'un grand nombre d'erreurs ou une performance correcte mais lente. Ces éventuelles dissociations pourront être analysées à la lumière des temps de réaction recueillis de manière objective par l'informatisation.
La vitesse de traitement est un indice de degré d'automatisation, c'est-à-dire du coût cognitif de l'activité. Plus le processus est rapide, automatisé, plus il est inconscient. Le sujet n'a pas conscience qu'il effectue une série d'opérations mentales qu'il ne contrôle pas volontairement et qui mobilisent très peu de ressources mentales, *Bonin, (2003)*. *Martin (1999)* note que l'ordinateur permet la conception de tâches spécifiques qui facilitent la mesure et l'interprétation des temps de réaction. La psychologie cognitive, qui s'intéresse notamment au temps de latence entre la présentation du stimulus et la réponse du sujet et au degré d'automatisation des procédures, s'inscrit dans cette perspective et offre un cadre théorique à l'interprétation des temps de réaction. En effet, *« les traitements cognitifs sont consommateurs de temps » et leur « dysfonctionnement devrait se traduire par un ralentissement des processus engagés », Lété (2004)*. Ces données peuvent aider le thérapeute à envisager les prochaines étapes de la thérapie et à objectiver les progrès de la thérapie. De plus, il faut tenir compte du biais que peut constituer le calcul des temps de réaction par des examinateurs différents, qui peuvent infléchir les résultats, *Le Gall et Allain (2001)*, un examinateur pouvant être plus réactif qu'un autre. Ce biais peut être contourné si la réponse du sujet est produite par lui-même soit par un système de bouton-réponse ou encore par un programme utilisant les réalités virtuelles.



...quelques situations concrètes

Tableau récapitulatif et ses légendes

Voici un exemple de tableau récapitulatif de normes obtenues pour un test donné ; **n** correspond au nombre de sujets ayant subi l'épreuve en vue de sa normalisation, **m** est la moyenne obtenue à l'épreuve et enfin α est l'écart type. L'écart type mesure la dispersion d'une série de valeurs autour de leur moyenne.
Pour calculer un écart type, il suffit de faire :

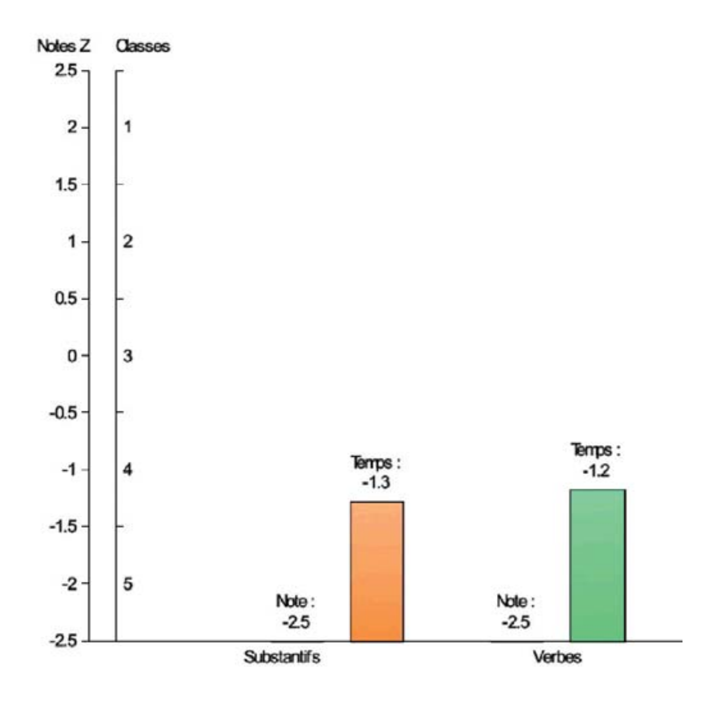
$$\frac{\text{Note brute obtenue par le patient} - \text{(moins) la moyenne}}{\text{Écart-type}}$$

• soit pour un enfant de 6 ans 2 mois ayant obtenu la note de 19 à l'épreuve ci après : l'écart- type réel de cet enfant sera :
19 – 24.10 / 4.44 = -1.14 écart-type (e.t) ou déviation standard

	Nombre de sujets	Score voc 1	
	N	m	α
5 ans 6	N= 153	12.67	3.52
6.0 ans	N= 158	24.10	4.44
6.6 ans	N= 155	31.34	8.24
7 ans 0	N= 158	35.25	8.19
7.6 ans	N= 156	40.72	8.50
8 ans 0	N= 158	44.18	6.20
8 ans 6	N= 152	52.26	6.39

Notes pondérées

Les notes brutes sont transformées en notes pondérées, qui permettent de situer l'enfant dans un étalonnage.
Prenons l'exemple de ce graphe :

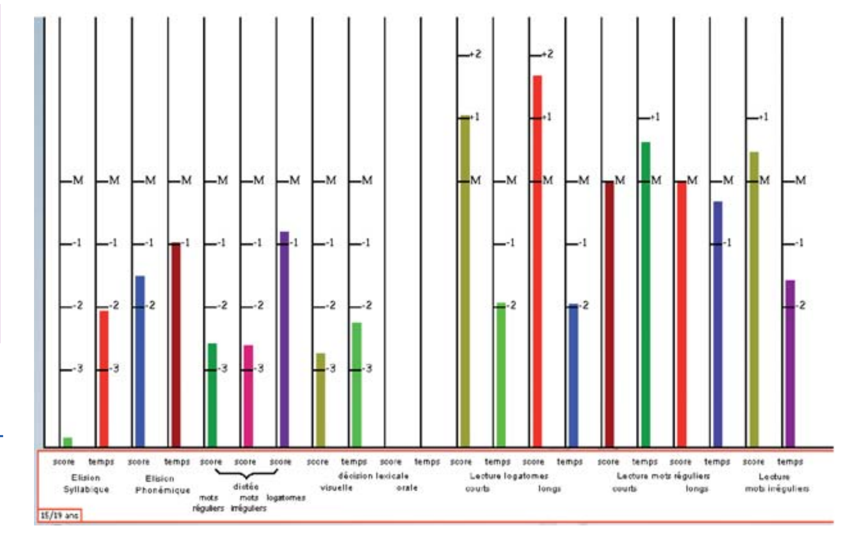


Pour les substantifs, ce patient obtient un score qui le situe à -2.5 écart-type (e.t) de la norme (notes Z), ce qui correspond à la note pondérée 5 de l'échelle des classes.
Généralement, que ce soit avec des logiciels de correction ou en fin de protocoles ou de manuel, les notes pondérées sont traduites en chiffres (1 à 5) qui correspondent :

1	2	3	4	5
Notes <-2 e.t	- 1.99<notes<-0.99	moyenne	0.99<notes<+1.99 e.t	Notes>+1.99 e.t

Le profil

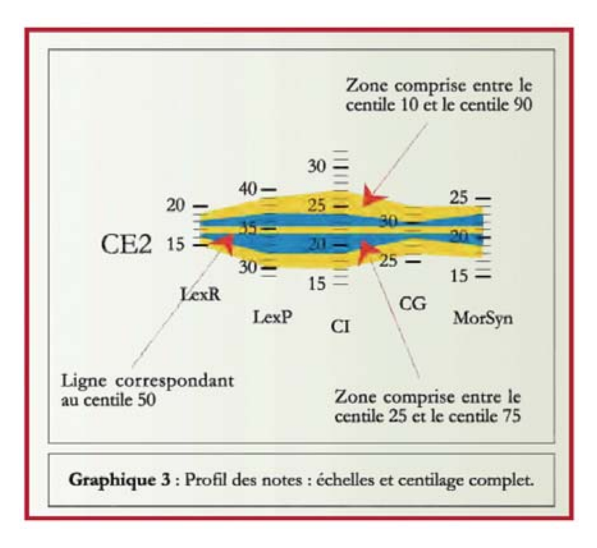
Il permet :
• de mesurer les compétences d'un sujet par rapport à sa classe d'âge, de niveau (interindividuelles),
• de voir où se situent les différences intra-individuelles, zones de contraintes (lorsque le profil est hétérogène).



Ici à nouveau les notes représentées correspondent à des notes pondérées.
Prenons ce patient, âgé de 19 ans, il se situe à au moins – 2 écart type de la moyenne (représentée par le M) pour l'épreuve de dictée de mots réguliers. (6e histogramme en rose sur le graphe).
Pour calculer l'écart type réel, il suffit de se reporter généralement à la fin du manuel du test ou dans la rubrique « consignes » s'il s'agit d'un logiciel pour reprendre le tableau récapitulatif de normes (cf a ci-dessus) . Pour ce cas présent la moyenne à l'épreuve de dictée de mots réguliers est de 17.8 et l'écart type de 1.47 soit : **14 (note brute obtenue - 17.8 (moyenne des sujets à l'épreuve) / par α 1.47 = -2.58. Ce patient se situe bien à – 2 écarts type de la moyenne.**

Les graphes

Lire un graphe, c'est avant tout s'assurer des valeurs, notes pondérées, comme ci-dessus ou percentiles comme ci-après. Un centile est chacune des 99 valeurs qui divisent les données triées en 100 parts égales.



Graphique 3 : Profil des notes : échelles et centilage complet.

Le centile 50 partie centrale en Jaune. Le centile 50 veut dire que 50 % des enfants obtiennent un résultat égal ou inférieur et que 50 % des enfants obtiennent un résultat supérieur.

Le centile 25 (parties en bleu) indique que 25 % des enfants obtiennent un résultat égal ou inférieur et que 75 % des enfants obtiennent un résultat supérieur.
Quant au Centile 75, il souligne résultat égal ou inférieur pour 75 % des enfants et un résultat supérieur pour seulement 25 % des enfants.

Enfin le centile 10 (parties externes en jaune) met en évidence que 10 % des enfants obtiennent un résultat égal ou inférieur mais que 90 % des enfants obtiennent un résultat supérieur.
Inversement le Centile 90 confirme que 90 % des enfants obtiennent un résultat égal ou inférieur et que 10 % des enfants obtiennent un résultat supérieur.

Conclusion

L'avancée des technologies, l'informatisation des outils d'évaluation rendent compte de la nécessité de considérer la variable temporelle comme élément diagnostique essentiel dans l'évaluation du langage. De ce fait, elle permet d'envisager de nouvelles pistes de réflexion pour la pratique évaluative et la prise en charge thérapeutique. Un traitement de l'amélioration conjointe de la qualité des productions et des temps de réponse lors de la résolution de tâches linguistiques sera à envisager dans chaque évaluation. Cette notion essentielle de ralentissement mental, à présent mesurable, nous paraît essentielle à considérer par les professionnels de la santé en vue d'une meilleure évaluation du handicap et d'éventuels ajustements soit au niveau scolaire soit du poste de travail.

Bibliographie

Bacher, F., Reuchlin, M. : *Les différences individuelles dans le développement cognitif de l'enfant,* Paris, PUF, 320 p. (1989).

Beech, J.R., Harding, L. et coll : *Tests, mode d'emploi, guide de psychométrie,* (trad. de J.-Luc Mogenet), Paris, ECPA, 180 p. (1994).

Binet, A., Simon, T. : *New methods for the diagnosis of the intellectual level of subnormals,* l'année psychologique, 12, 191-244. (1905).

Bonin, P. : *Production verbale de mots, Approche cognitive.* De Boeck Université, Bruxelles. (2003).

Cibois, P. : *L'analyse factorielle,* Paris, PUF, 128 p. (1983).

Dickes, P., Tournois, J., Flieller, A., Kopj, L : *La psychométrie,* Paris, PUF, 288 p. (1994).

French, C. : *L'évaluation assistée par ordinateur, Chapitre 7, in : Tests, mode d'emploi, Guide de psychométrie, sous la direction de J.R Beech et L. Harding, Paris :* Centre de psychologie Appliquée, 159-67, 180 p. (1994).

Gatignol, P. : *Evaluations et bilans : le point de vue de l'orthophoniste. Intérêt d'une évaluation spécifique en vue d'une rééducation ciblée.* Les Actes des 5es journées scientifiques de l'Ecole d'Orthophonie de Lyon : 25-29. (2004b).

Gatignol, P., Duffau, H., Plaza, M. : *Influence de la variable temporelle sur les performances d'accès au lexique.*

Huteau, M., Lautrey, J. : *Les tests d'intelligence,* Paris, La Découverte, Repères 229, 123 p. (1997).

Huteau, M., Lautrey, J. : *Evaluer l'intelligence, Psychométrie cognitive,* Paris, PUF, 310 p. (1999).

Kim, J.O., Mueller, C.W. : *Introduction to factor analysis: what it is and how to do it,* Beverly Hills CA, Sage Publications, 79 p. (1978).

Le Gall, A., Allain, P. : *Applications des techniques de réalité virtuelle à la neuropsychologie clinique. Champ psychosomatique,* L'Esprit du temps, 2001 ; 22(2) :25-38,170 p.(2001).

Lété B. : *La chronométrie mentale appliquée à l'évaluation diagnostic de la lecture.* Les actes de 3e journée scientifique de l'école d'orthophonie de Lyon : Bilans et évaluation en orthophonie, UCLB., pp. 81-83. (2004).

Metz-Lutz M.N. : *Les tests dans le bilan d'aphasie : intérêt diagnostique, thérapeutique et heuristique,* Bulletin Audiophonologie Ann. Sc. Université de Franche-Comté, 4, 3, 287-300. (1988).

Oudry, M., Gatignol, P., Robert, A.M, Plaza, M. : *Création et Validation d'un Bilan Informatisé de Langage Ecrit chez l'adolescent et l'adulte.*

Pichot, P. : Les Tests Mentaux, Paris, PUF, *Que Sais-je* n° 626, 128 p. (1997).

Reuchlin, M. : *La mesure en psychologie,* in Fraisse P., Piaget J. (éds), Traité de psychologie expérimentale, 3e éd., Paris, PUF, Vol 1, 207 p. (1970).

Reuchlin, M. : *Précis de statistiques,* 7e éd., Paris, PUF, 1998, 256 p. (1998).

Stevens, S.S. : *Mathematics, measurement and psychophysics,* in Stevens S.S. (ed.), Handbook of experimental psychology, New York, Wiley, pp 1-49, 1436 p. (1951).

Tran, T.H., Duquenne, J., Moreau, E. : *Les troubles de la dénomination. Déficits et stratégie. Proposition d'une grille d'analyse des réponses obtenues en dénomination d'images.* Glossa 71 : 4-16. (2000).

Vrignaud, P. : *Les tests au XXIe siècle. Que peut-on attendre des évolutions méthodologiques et technologiques dans le domaine de l'évaluation psychologique des personnes ?* Pratiques psychologiques, 4, 5-27. (1996).

Vrignaud, P. : *Aspects théoriques et méthodologiques généraux liés à l'évaluation : l'exemple de l'évaluation de la lecture,* 5e journées d'Orthophonie, Lyon, UCLB, 7-15. (2004).

Vrignaud, P., Castro, D., Mogenet, J.-L : *Recommandations Internationales sur l'Utilisation des Tests,* version française élaborée pour la Société Française de Psychologie, Pratiques Psychologiques, L'esprit du temps, hors-série juin, 33 p. (2003).

Walton, R., Bartram, M. : *PET : Preliminary English Test, Teacher's resource book,* Walton on Thames, Nelson 158 p. (1994).